Moura, P. L., Dobbe, J. G. G., Streekstra, G. J., Rab, M. A. E., Veldthuis, M., Fermo, E., van Wijk, R., van Zwieten, R., Bianchi, P., Toye, A. M., & Satchwell, T. (2020). Rapid diagnosis of hereditary haemolytic anaemias using automated rheoscopy and supervised machine learning. *British Journal of Haematology*. https://doi.org/10.1111/bjh.16868

Rapid diagnosis of hereditary haemolytic anaemias using automated rheoscopy and supervised machine learning

Pedro L. Moura[1,2#], Johannes G.G. Dobbe[3], Geert J. Streekstra[3], Minke A.E. Rab[4,5], Martijn Veldthuis[6], Elisa Fermo[7], Richard van Wijk[4], Rob van Zwieten[6,8], Paola Bianchi[7], Ashley M. Toye[1,2,9*], Timothy J. Satchwell[1,2,9*]

[1]*School of Biochemistry, University of Bristol, UK*

[2]*NIHR Blood and Transplant Research Unit in Red Cell Products, University of Bristol, UK*

[#]*Present address: Center for Hematology and Regenerative Medicine, Karolinska Institutet, Department of Medicine, Karolinska University Hospital, Huddinge, Stockholm, Sweden*

[3]*Amsterdam UMC, University of Amsterdam, Department of Biomedical Engineering and Physics, Meibergdreef 9, Amsterdam, The Netherlands*

[4]*Department of Clinical Chemistry and Haematology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*

[5]*Van Creveldkliniek, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*

[6]*Department of Blood Cell Research, Sanquin Amsterdam, Amsterdam, The Netherlands*

[7]*UOC Ematologia, UOS Fisiopatologia delle Anemie, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano, Italy*

[8]*Laboratory of Red Blood Cell Diagnostics, Sanquin Amsterdam, Amsterdam, The Netherlands*

[9]*Bristol Institute for Transfusion Sciences, National Health Service Blood and Transplant (NHSBT), UK*

**These authors contributed equally to this work*

Corresponding authors: t.satchwell@bristol.ac.uk, ash.m.toye@bristol.ac.uk

Main text word count: 982 words

Introduction

Haemolytic anaemias arise when red blood cell (RBC) integrity is compromised, eventually resulting in premature clearance or lysis and leading to anaemia when these effects cannot be sufficiently compensated by the capacity of the bone marrow to produce new cells.(Ucar 2002) Hereditary anaemia occurs as a consequence of genetic mutation(Risinger*, et al* 2019) (e.g. affecting membrane complex or cytoskeletal proteins, haemoglobin or metabolic enzymes), and diagnosing affected patients is a complex process since, given the wide variety of possible genetic causes, multiple examinations must be performed and an unambiguous result is usually reached only after DNA sequencing.(Kim*, et al* 2017) Furthermore, phenotypic severity can vary widely not just among individuals with different mutations but among individuals suffering from the same mutation, thereby complicating diagnosis.(Glogowska*, et al* 2017)

While molecular diagnoses have become increasingly easier, cheaper and faster to perform in recent years, constraints on their use still exist,(Di Resta*, et al* 2018) and phenotype-based diagnostic methods still constitute an important proposition. Ektacytometry is a standard diagnostic platform for RBC disorders(Da Costa*, et al* 2016, Johnson and Ravindranath 1996) but only provides cell population-based data and requires a trained expert for data interpretation. Single cell rheoscopy can provide additional information, with higher complexity as a drawback; however, analysis of such data could potentially be facilitated by the use of machine learning (ML; automated, algorithm-based systems that generate data-driven predictions(Nichols*, et al* 2019)).

We present here a preliminary framework for automated rheoscopy-based diagnosis of several types of hereditary haemolytic anaemia samples (**Figure 1A**) that requires low sample volumes and is efficient, rapid and expandable.

Methods

Peripheral Blood Donor and Patient Samples

Healthy control donor and diagnosed patient samples were collected according to procedures approved by research ethics committee and in accordance with the Declaration of Helsinki. 47 blood samples were analysed at the University of Bristol following shipment from clinics in Milan or Utrecht (6 controls, 13 Hereditary Spherocytosis [HS] patients, 9 Congenital Dyserythropoietic Anaemia type II [CDAII] patients, 6 Pyruvate Kinase Deficiency [PKD] patients, 10 Hereditary Xerocytosis / Dehydrated Hereditary Stomatocytosis (DHS) 1 [HX] patients and 3 Gardos Xerocytosis / DHS2 [GX] patients). A further 26 samples were analysed at Sanquin (11 controls, 7 HS patients and 8 Hereditary Elliptocytosis [HE] patients).

Automated Rheoscope and Cell Analyzer

1 µL of whole blood was diluted in 200 µL of a polyvinylpyrrolidone solution (viscosity 28.1 mPa·s). Samples were assessed in an Automated Rheoscope and Cell Analyzer (ARCA) according to published protocols.(Moura*, et al* 2018) At least 1000 cells per sample were analysed, providing deformability index (DI) and cross-sectional area (area) quantification.

Computational analysis

A Python script was developed for statistical analysis, data visualization and automatic dataset classification (**Data availability**). The full datasets used for training purposes were sampled and randomized into testing (500 cells) and training datasets (remainder). DI and area were normalized by the maximum measurable values (3.3/5.0 DI from Bristol and Sanquin, respectively, and 140 μm^2 area) and the training datasets were repeatedly subjected to random sampling to generate 10,000 subsets of 500 cells each, followed by calculation of the average and standard deviation of the DI and area. Each sample category was then attributed unique identifiers. Classifiers were generated with the scikit-learn package(Pedregosa*, et al* 2011), trained with the generated subsets and tested with the initial testing subsets. Classification of unseen datasets was performed by selecting the mode of the machine-selected identifiers after 10,000 classifications.

Results and Discussion

We have demonstrated in previous work that automated rheoscopy-based analyses can elucidate differences arising from reticulocyte maturation(Moura*, et al* 2018) as well as loss of cellular stability.(Moura*, et al* 2019) A particularly interesting observation from the same work was the fact that combining single-cell deformability index (DI) and cross-sectional area measurements provides a novel metric (**Figure 1B**) which to date has not been examined in the context of disease diagnosis.

Therefore, we evaluated whole blood samples from diagnosed anaemic patients of varied aetiologies (HS, CDAII, PKD, HX and GX) against healthy donors using the proposed methodology (**Figure 1C**). Crucially, despite these diseases being frequently misdiagnosed due to overlapping clinical or morphological phenotypes,(Danise*, et al* 2001, Fermo*, et al* 2017) we observed them to display unique rheoscopy “fingerprints” upon visualization.

Machine learning algorithms were next explored to automate the classification of ARCA data and thus facilitate the processing of larger numbers of samples, A flow chart listing the procedure used for these attempts is displayed in **Figure 2A**.

To provide sufficient information for training a ML classifier, the data was augmented through random sampling, vastly extending the number of new datasets with similar characteristics. We then tested the trained classifiers on a combination of fully unseen data and the testing sets generated before augmentation. A full summary of the prediction accuracies achieved (and listing the best performing classifiers) is provided in **Figure 2B** with the best performing algorithm correctly identifying sample datasets with 92% accuracy. (**Figure 2C**). We note that the GX samples were excluded due to the sample number being too low for classifier training.

For further verification, the classifiers were retrained on additional samples (11 controls, 7 HS patients and 8 HE patients) obtained on a second ARCA device in an independent laboratory and using different acquisition settings. Again, we observed increasing classification accuracy until the

use of 6 training datasets (at which point the classifier likely overfits these data), as per **Figure 2D**, achieving a final prediction accuracy for multiclass classification that is comparable to that offered by osmotic gradient ektacytometry when classifying HS samples alone.(Llaudet-Planas*, et al* 2018)

Importantly, the best-performing algorithms utilized here achieve complete differentiation between controls and diseased patients and accurately identify a variety of disorders potentially allow for the rapid preliminary identification or discrimination of more elusive diseases(Zaninoni*, et al* 2018) (such as CDAII and PKD) without time-consuming laboratory assays or molecular testing methods. Furthermore, the possibility to continuously incorporate data from new samples or the expansion with haematological conditions beyond those characterized in this study may ultimately allow for diagnosing a large number of samples in a relatively short period using minimal sample volumes. In conclusion, the method described in this work represents an exciting step forward towards facilitating the improved diagnosis of haemolytic anaemias.

Data availability

All raw ARCA datasets obtained during this study, Python scripts generated for dataset analysis, classifier training and sample classification and the confusion matrices generated for classifier evaluation have been made publicly available through the following Github repository: https://github.com/pedrolmoura/ARCA-ML.

Acknowledgments

The authors would like to thank the donors, patients and their family members for their willingness to participate in this research. The authors thank Mr. Ario Sadafi (Technische Universität München, Munich, Germany) for helpful discussions regarding feature extraction for machine learning algorithm development. PLM was funded by the European Union (H2020-MSCAITN-2015, “RELEVANCE”, Grant agreement number 675117). MAER is supported by the Eurostars grant estar18105 and by an unrestricted grant provided by RR Mechatronics. PB was funded by the Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Grant no. 2019 175/02, 2019. AMT and TJS were funded/supported by an NHS Blood and Transplant (NHSBT) R&D grant (WP15-05) and the National Institute for Health Research Blood and Transplant Research Unit (NIHR BTRU) in Red Cell Products (IS-BTU-1214-10032). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Authorship

Contribution: PLM acquired data, prepared figures and developed the Python code for dataset analysis and classification. JGGD and GJS provided essential ARCA equipment and image analysis software. MAER and RvW diagnosed HX patients and provided blood samples. MV and RvZ diagnosed HS and HE patients and performed initial ARCA analysis. EF and PB diagnosed HS, CDAII, HX and PKD patients and provided blood samples. PLM, AMT and TJS conceived and designed experiments and wrote the manuscript. TJS and AMT contributed equally to conception and supervision of the work. All authors read and edited the manuscript.

Conflict of interest statement: The authors declare no competing financial interests.

Correspondence: Timothy J. Satchwell, School of Biochemistry, University of Bristol, Medical Science Building, University Walk, Bristol BS8 1TD, United Kingdom; e-mail: t.satchwell@bristol.ac.uk; or Ashley M. Toye, School of Biochemistry, University of Bristol, Medical Science Building, University Walk, Bristol BS8 1TD, United Kingdom; e-mail: ash.m.toye@bristol.ac.uk.

References

Da Costa, L., Suner, L., Galimand, J., Bonnel, A., Pascreau, T., Couque, N., Fenneteau, O., Mohandas, N., Society of, H., Pediatric Immunology, g. & French Society of, H. (2016) Diagnostic tool for red blood cell membrane disorders: Assessment of a new generation ektacytometer. *Blood Cells Mol Dis,* **56,** 9-22.

Danise, P., Amendola, G., Nobili, B., Perrotta, S., Miraglia Del Giudice, E., Matarese, S.M., Iolascon, A. & Brugnara, C. (2001) Flow-cytometric analysis of erythrocytes and reticulocytes in congenital dyserythropoietic anaemia type II (CDA II): value in differential diagnosis with hereditary spherocytosis. *Clin Lab Haematol,* **23,** 7-13.

Di Resta, C., Galbiati, S., Carrera, P. & Ferrari, M. (2018) Next-generation sequencing approach for the diagnosis of human diseases: open challenges and new opportunities. *EJIFCC,* **29,** 4-14.

Fermo, E., Vercellati, C., Marcello, A.P., Zaninoni, A., van Wijk, R., Mirra, N., Curcio, C., Cortelezzi, A., Zanella, A., Barcellini, W. & Bianchi, P. (2017) Hereditary Xerocytosis due to Mutations in PIEZO1 Gene Associated with Heterozygous Pyruvate Kinase Deficiency and Beta-Thalassemia Trait in Two Unrelated Families. *Case Rep Hematol,* **2017,** 2769570.

Glogowska, E., Schneider, E.R., Maksimova, Y., Schulz, V.P., Lezon-Geyda, K., Wu, J., Radhakrishnan, K., Keel, S.B., Mahoney, D., Freidmann, A.M., Altura, R.A., Gracheva, E.O., Bagriantsev, S.N., Kalfa, T.A. & Gallagher, P.G. (2017) Novel mechanisms of PIEZO1 dysfunction in hereditary xerocytosis. *Blood,* **130,** 1845-1856.

Johnson, R.M. & Ravindranath, Y. (1996) Osmotic scan ektacytometry in clinical diagnosis. *J Pediatr Hematol Oncol,* **18,** 122-129.

Kim, Y., Park, J. & Kim, M. (2017) Diagnostic approaches for inherited hemolytic anemia in the genetic era. *Blood Res,* **52,** 84-94.

Llaudet-Planas, E., Vives-Corrons, J.L., Rizzuto, V., Gomez-Ramirez, P., Sevilla Navarro, J., Coll Sibina, M.T., Garcia-Bernal, M., Ruiz Llobet, A., Badell, I., Velasco-Puyo, P., Dapena, J.L. & Manu-Pereira, M.M. (2018) Osmotic gradient ektacytometry: A valuable screening test for hereditary spherocytosis and other red blood cell membrane disorders. *Int J Lab Hematol,* **40,** 94-102.

Moura, P.L., Hawley, B.R., Dobbe, J.G.G., Streekstra, G.J., Rab, M.A.E., Bianchi, P., van Wijk, R., Toye, A.M. & Satchwell, T.J. (2019) PIEZO1 gain-of-function mutations delay reticulocyte maturation in hereditary xerocytosis. *Haematologica*.

Moura, P.L., Hawley, B.R., Mankelow, T.J., Griffiths, R.E., Dobbe, J.G.G., Streekstra, G.J., Anstee, D.J., Satchwell, T.J. & Toye, A.M. (2018) Non-muscle myosin II drives vesicle loss during human reticulocyte maturation. *Haematologica,* **103,** 1997-2007.

Nichols, J.A., Herbert Chan, H.W. & Baker, M.A.B. (2019) Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophys Rev,* **11,** 111-118.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research,* **12,** 2825-2830.

Risinger, M., Emberesh, M. & Kalfa, T.A. (2019) Rare Hereditary Hemolytic Anemias: Diagnostic Approach and Considerations in Management. *Hematol Oncol Clin North Am,* **33,** 373-392.

Ucar, K. (2002) Clinical presentation and management of hemolytic anemias. *Oncology (Williston Park),* **16,** 163-170.

Zaninoni, A., Fermo, E., Vercellati, C., Consonni, D., Marcello, A.P., Zanella, A., Cortelezzi, A., Barcellini, W. & Bianchi, P. (2018) Use of Laser Assisted Optical Rotational Cell Analyzer (LoRRca MaxSis) in the Diagnosis of RBC Membrane Disorders, Enzyme Defects, and Congenital Dyserythropoietic Anemias: A Monocentric Study on 202 Patients. *Front Physiol,* **9,** 451.

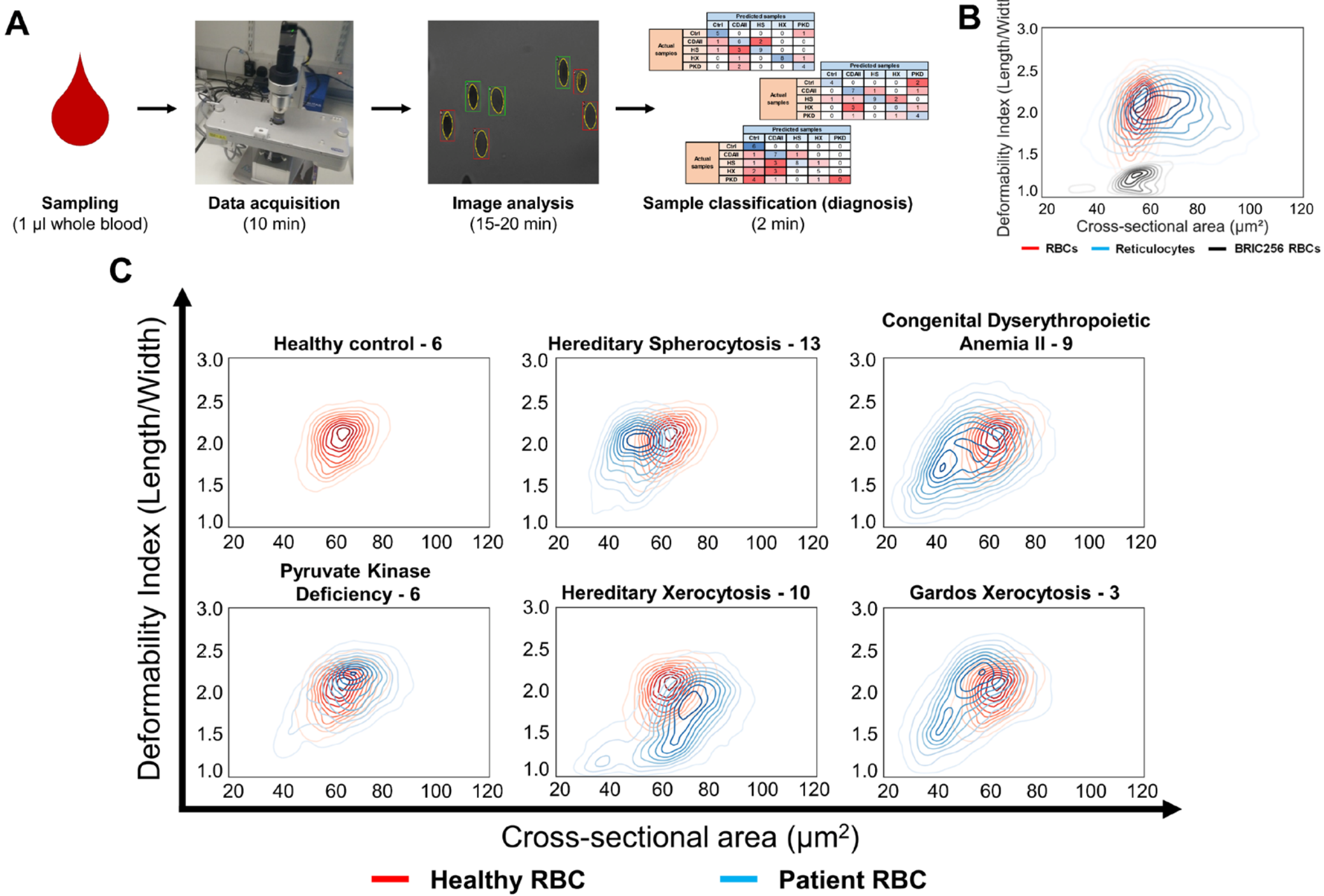
Figure legends

Figure 1 – Different hereditary rare anaemias display distinct area and deformability profiles
A) Design of the method for automatic sample classification. Whole blood is collected by the clinician, and a sample is obtained and processed using an Automated Rheoscope and Cell Analyzer (ARCA). Images acquired are subjected to computational analysis to determine cross-sectional area and deformability of at least 1000 individual cells, and the resulting datasets are then classified through trained computational models, achieving a diagnosis in less than 30 minutes.
B) Contour plots of cross-sectional area plotted against the deformability index (as measured by dividing cell length by cell width), visualizing the probability distribution of erythrocytes (RBCs), cultured reticulocytes (reticulocytes) and erythrocytes treated with an anti-Glycophorin A antibody (BRIC256, International Blood Group Reference Laboratory) before analysis to induce membrane stiffening (BRIC256 RBCs). The control erythrocyte and cultured reticulocyte data shown in this panel were previously reported in Moura et al., 2018.(Moura*, et al* 2018) A minimum of 1000 cells were analysed per sample. All samples were analysed using the ARCA.
C) Contour plots of cross-sectional area plotted against the deformability index (as measured by dividing cell length by cell width), visualizing the probability distribution of patient samples overlaid to allow for comparison with healthy controls. A minimum of 1000 cells were analysed per blood sample. All samples were analysed using the ARCA. The samples are listed from left to right:
Top row: healthy controls (n = 6), Hereditary Spherocytosis patients (n = 13), Congenital Dyserythropoietic Anaemia II patients (n = 9).
Bottom row: Pyruvate Kinase Deficiency patients (n = 6), Dehydrated Stomatocytosis type 1 or Hereditary Xerocytosis patients (n = 10), Dehydrated Stomatocytosis type 2 or Gardos Xerocytosis patients (n = 3).

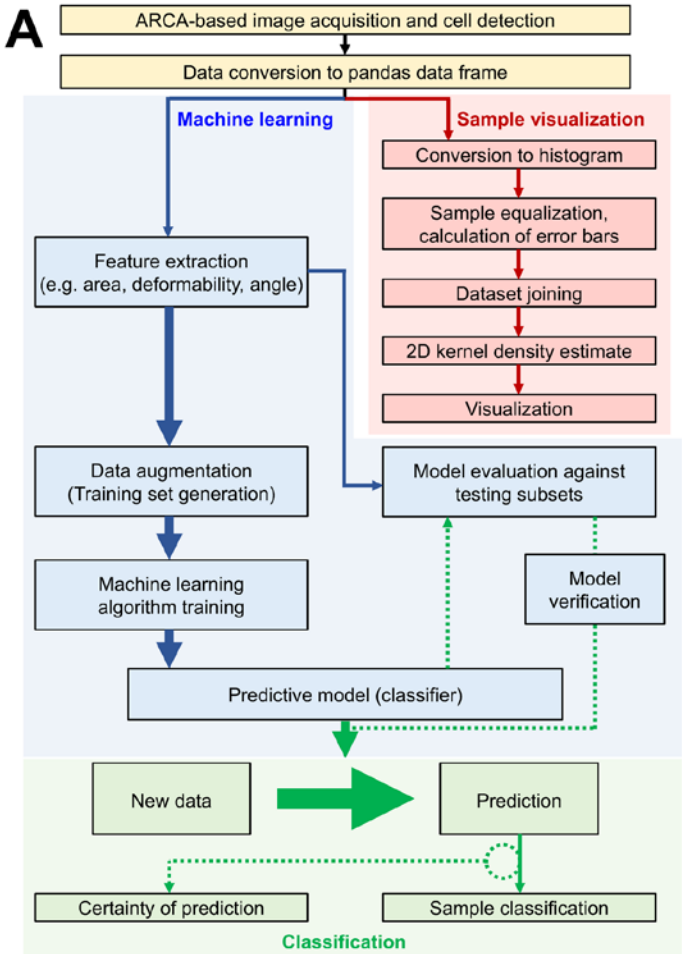
Figure 2 – Machine learning-based classification of automated rheoscopy datasets provides accurate diagnoses for unseen samples
A) Flow diagram outlining the procedure for ARCA-based data visualization and automated sample classification. The sample is first analysed to produce a raw data table. These data are then reorganized into a Python pandas (“panel data”) data frame for ease of processing. If visualization is required, samples from a given sample type are equalized in cell number, joined and subjected to kernel density estimation to estimate the probability density functions of analysed features (e.g. cross-sectional area, deformability index, cell angle) and then visualized through contour plots or scatter plots. Data to be used for machine learning undergo feature extraction (removal of all non-essential information) and a subsection is sampled randomly (without reposition) for creation of a testing set. The remaining data then undergoes augmentation by generation of a series of randomly sampled datasets (with reposition, 10,000x) which will be used for training a supervised machine learning algorithm. After training, a predictive model (i.e. classifier) is generated which first is tested with the previously generated testing set. Upon satisfactory results with the testing set, the classifier can then generate predictions for new unseen data. The final results consist of a sample label (or classification) and the certainty of that classification through repeated exposure to the same sample.
B) Comparison of the overall prediction accuracy of multiple supervised machine learning algorithms in ARCA-based automated sample diagnosis as a function of the number of datasets per condition used for classifier training (from 0 datasets used, which should result in a random diagnosis, to a maximum of 6 datasets), comparing the samples analysed at the University of Bristol (except GX samples, which were too few to analyse). Prediction accuracy is coloured on a percentage scale from red (0%) to blue (100%). The best-performing algorithm per no. of datasets is bolded in the accuracy matrix. The graph displays the average prediction accuracy of all algorithms (blue). Error bars = ±SD. The prediction accuracies of the best-performing algorithms are plotted in green, while the prediction accuracies of the worst-performing algorithms are plotted in red.
C) Prediction accuracy of the best performing algorithm in **B)**. The samples used consist of healthy controls, congenital dyserythropoietic anaemia II patients (CDAII), hereditary spherocytosis patients

(HS), hereditary xerocytosis patients (HX) and pyruvate kinase deficiency patients (PKD). Rows identify real samples provided, whilst columns identify the algorithm's prediction of the provided samples' identity. The blue diagonal indicates samples that were correctly diagnosed (true positives). Red cells in the surrounding matrix indicate incorrect diagnoses (i.e. 2 HS samples were misdiagnosed as CDAII and 1 HX sample was misdiagnosed as HS). Accuracy is provided as a percentage of the true positives within the total number of samples and is coloured on a percentage scale from red (0%) to blue (100%). Average accuracy is provided as an average of the accuracies for all sample types. Data for all other algorithms and sample numbers tested is provided in **Sup. Figures 1** to **7.**

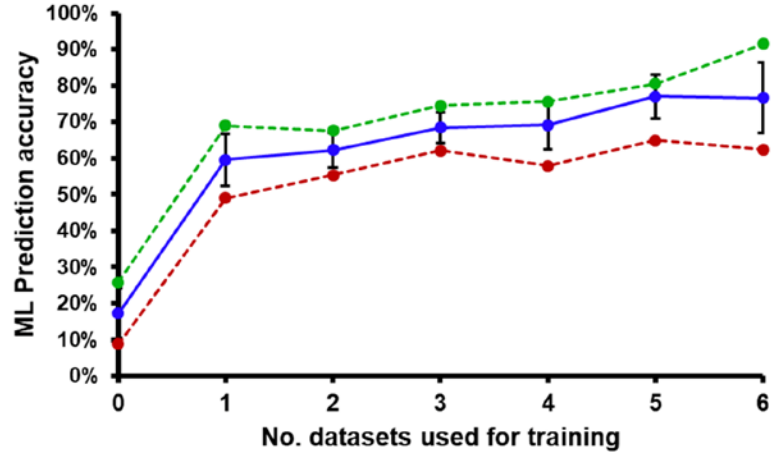
D) Comparison of the overall prediction accuracy of multiple supervised machine learning algorithms in ARCA-based automated sample diagnosis as a function of the number of datasets used for training, comparing samples from healthy controls (Ctrl), hereditary spherocytosis patients (HS) and hereditary elliptocytosis patients (HE) analysed at Sanquin. The graph displays the average prediction accuracy of all algorithms (blue). Error bars = ±SD. The prediction accuracies of the best-performing algorithms are plotted in green, while the prediction accuracies of the worst-performing algorithms are plotted in red.


A
Sampling
(1 µl whole blood)
Data acquisition
(10 min)
Image analysis
(15-20 min)
Sample classification (diagnosis)
(2 min)
Predicted samples
Actuel samples
Ctrl
CDAII
HS
HX
PKD
B
Deformability Index (Length/Width)
Cross-sectional area (µm²)
RBCs
Reticulocytes
BRIC256 RBCs
C
Healthy control - 6
Hereditary Spherocytosis - 13
Congenital Dyserythropoietic Anemia II - 9
Pyruvate Kinase Deficiency - 6
Hereditary Xerocytosis - 10
Gardos Xerocytosis - 3
Deformability Index (Length/Width)
Cross-sectional area (µm²)
Healthy RBC
Patient RBC

A



B



Prediction accuracy	No. Datasets used for training						
	0	1	2	3	4	5	6
K-Nearest Neighbours Classifier	20.00%	62.34%	66.51%	69.85%	73.18%	79.85%	75.67%
Bagging Classifier	**25.90%**	54.38%	61.40%	**74.55%**	**75.62%**	78.31%	**91.54%**
AdaBoost (SAMME)	20.00%	49.03%	64.02%	66.74%	57.86%	64.91%	62.36%
Multi-layer Perceptron Classifier	9.06%	**68.96%**	55.38%	70.35%	68.07%	80.07%	75.67%
Gradient Boosting	20.00%	58.29%	**67.59%**	66.74%	73.40%	**80.53%**	81.94%
Random Forests	8.87%	64.51%	58.26%	62.12%	66.31%	78.31%	72.34%
Minimum accuracy	8.87%	49.03%	55.38%	62.12%	57.86%	64.91%	62.36%
Average accuracy	17.30%	59.58%	62.19%	68.39%	69.07%	76.99%	76.59%
Standard deviation	6.85%	7.21%	4.77%	4.21%	6.52%	5.99%	9.74%
Maximum accuracy	25.90%	68.96%	67.59%	74.55%	75.62%	80.53%	91.54%

C

Bagging classifier

		Predicted samples						
		Ctrl	CDAll	HS	HX	PKD	Total	Acc
Actual samples	Ctrl	6	0	0	0	0	6	100%
	CDAll	0	9	0	0	0	9	100%
	HS	0	2	11	0	0	13	85%
	HX	0	0	1	9	0	10	90%
	PKD	0	0	0	0	6	6	100%
						Average accuracy		**92%**

D